INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Fundamental frequency accommodation in multi-party human-robot game interactions: The effect of winning or losing

Omnia Ibrahim[1], Gabriel Skantze[2], Sabine Stoll[3], Volker Dellwo[4]

[1]URPP Language and Space, University of Zurich, Switzerland
[2]Department of Speech Music and Hearing, KTH Royal Institute of Technology, Sweden
[3]Department of comparative linguistics, University of Zurich, Switzerland
[4]Department of computational linguistics, University of Zurich, Switzerland

omnia.ibrahim@spur.uzh.ch, skantze@kth.se, sabine.stoll@uzh.ch, volker.dellwo@uzh.ch

Abstract

In human-human interactions, the situational context plays a large role in the degree of speakers' accommodation. In this paper, we investigate whether the degree of accommodation in a human-robot computer game is affected by (a) the duration of the interaction and (b) the success of the players in the game. 30 teams of two players played two card games with a conversational robot in which they had to find a correct order of five cards. After game 1, the players received the result of the game on a success scale from 1 (lowest success) to 5 (highest). Speakers' f_o accommodation was measured as the Euclidean distance between the human speakers and each human and the robot. Results revealed that (a) the duration of the game had no influence on the degree of f_o accommodation and (b) the result of Game 1 correlated with the degree of f_o accommodation in Game 2 (higher success equals lower Euclidean distance). We argue that game success is most likely considered as a sign of the success of players' cooperation during the discussion, which leads to a higher accommodation behavior in speech.

Index Terms: Vocal accommodation, Fundamental frequency, Human-robot interaction, Multi-party interactions

1. Introduction

When people engage in speech communication, they tend to align their vocal characteristics with those of their interlocutor. This phenomenon is typically referred to as accommodation or convergence [1]. In the acoustic domain, accommodation has been demonstrated through multiple variables, such as speaking rate [2], intensity [3], fundamental frequency of oscillation (f_o) [4] [5] [6], or the interaction between those features [7]. In the present paper, we studied accommodation in terms of f_o. We focused on a novel situation, in which two humans communicate with a social robot to solve a game task (henceforth: multi-party human-robot game interaction). We were interested in two questions:

(a) To what degree do humans accommodate to each other (H-H accommodation) and to what degree to the social robot (H-R accommodation) in terms of f_o?

(b) What are the factors driving accommodation behavior in H-H and H-R interaction?

According to the communication accommodation theory [8], humans typically change their vocal characteristics to align with their interlocutors in face-to-face conversations. Such changes can lead to changes in speaking style in some cases. The theory claims that convergence between conversational partners is an intentional process and somehow predicted over the course of the conversation [8]. Numerous studies viewed convergence as a default and sometime uncontrolled behavior during the conversation [9] [10] [11]. Lewandowski [12] found that native English speakers still converge toward their native German-speaking interlocutors' accents even though they have been explicitly instructed not to change their pronunciation to accommodate to their interlocutors' non-native accents. Moreover, Brennan and Clark [13] discovered that speakers design their speech specifically for their conversational partners, and they adapt to their interlocutor's new conceptualization of objects over the course of a conversation [14].

There have been several studies, which investigate the factors that drive accommodation. There has been first evidence that accommodation increases with the duration of an interaction [15]. Episodic models of speech production [16] could propose an explanation to the speakers' ability to accommodate over the course of the conversation [15]; when we perceive an instance of a particular category, it becomes part of the definition of that category; and the following productions of instances of that category are influenced by the category's new definition [16]. Other studies have questioned this view of accommodation phenomena. Bane et al. [17] showed in their study of voice onset time accommodation that the convergence in the speakers' vocal characteristics is not in a single direction throughout the analyzed duration. This suggests that accommodation might not be linear over time, and is conditioned by the social roles of the speakers. One potential reason for this variation in the degree of accommodation is due to social factors; such as social characteristics of the speakers and the relationship between the interlocutors are significant predictors of convergence [18]. Relevant factors include gender [18], dialect [8] [19], interlocutor status [17], and attitude towards a model talker [11]. Situational factors (such as effects of the conversational topic or task) also contribute to the degree of accommodations between speakers in human-human and human-machine interactions [20] [21].

In the present study, one particular social bonding factor was present in the experiment, which was the success of the humans in a card game during an interaction with a social robot. We investigated whether this success/non-success could impact the way H-H and H-R accommodation evolved. It seems likely that humans accommodate more to each other when they win as a team as opposed to when they lose. It is

http://dx.doi.org/10.21437/Interspeech.2019-2496

also plausible that the interaction with the social robot becomes more involved when winning a game, which could result in stronger accommodation to the robot.

It has been reported that the accommodation in f_o is more frequent in conversations than the accommodation of other prosodic parameters [22]. Furthermore, listeners typically adapt more rapidly to their conversational partners' f_o [6]. For example, they can identify the location of f_o values relative to an individual speaker's range [23]. This may be a result of listeners' expectations regarding average f_o for different genders [24]. It has also been found that f_o accommodation has a conversational function in turn-taking in overlapping speech [25] [6] [22]. As such, f_o is a crucial interactive conversation structuring parameter. In their recent work [4], the relation between f_o accommodation and turn-taking have been investigated using the same two approaches previously proposed by Schweitzer [6], f_o initialising (local context with adjacent turns) and f_o normalizing (model of other speaker' f_o norms). They found that f_o accommodation is only relevant as a turn competitive resource in overlaps that start clearly before a speaker transition. Their results suggested that both f_o initialisation and normalisation take place when speakers compete for the turn in overlap.

Based on the literature, in this paper, we investigate the f_o accommodation in the H-H-R game playing setting. We hypothesized that social factors (e.g. players' gender) and situational factors (e.g. game result) along with the duration of the interaction will influence the degree of accommodation between H-H and H-R. Our aim was to explore to what degree these factors can play a role in H-H-R interaction.

2. Data and Methodology

The data are extracted from multi-party human-robot discussion corpus [26], which were collected at an exhibition in the Swedish National Museum of Science and Technology for nine days. The Swedish corpus consists of conversations that were recorded during collaborative card games with a social robot (Furhat).

2.1. Participants

There were 60 adult participants (30 males and 30 females) who played several games with the robot (Furhat) in paired teams. There were 30 teams: 10 teams were male – male teams, 10 female - female and 10 male - female teams. The age range of male players was 16-64 with a mean of 35 ages, while female players ranged from 16-80 years with a mean of 37.

2.2. Recording setup

The interactional setting of the game is illustrated in Figure 1. Two players were seated at a large table with a multi-touch screen, opposite the Furhat robot head (see Figure 1), which has an animated face back-projected on a translucent mask, as well as a mechanical pan tilt neck [27]. This allows Furhat to direct the gaze using a combination of head and eye movements. The animated face allows for very accurate and expressive lip movements, facial gestures and gaze, which have been shown to be easy for users to read [27].

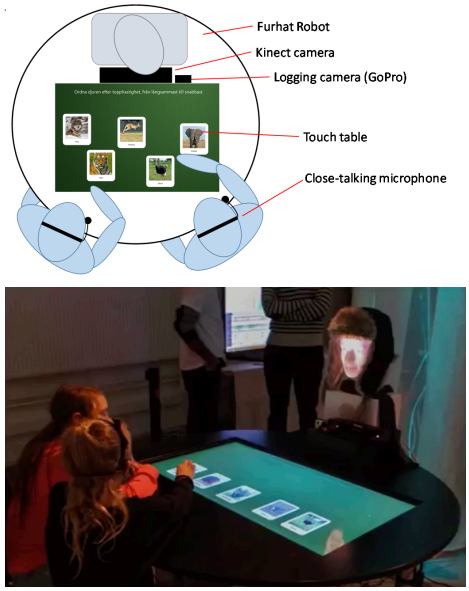

Figure 1: *The setup used in the museum*

The synthetic voice is also complemented by non-verbal expressions, such as sighs, breathing, filled pauses and different types of backchannels. Both users were wearing unidirectional headset microphones, which allowed for the recording of two separate good quality audio streams (given the noisy setting in the museum). The speech to noise ratio in the recording is ≈ 38 dB. A Kinect camera was used to track the location and rotation of the users' heads.

2.3. Procedure

The team was seated at a table and the recordings started when they pressed a 'Start' button on the touch screen. Furhat (Robot) initiated the interaction by asking them for their names. Then five cards are shown on the table and Furhat explained the game, which consists of sorting 5 cards according to sorting criterion, after which the discussion starts. An example interaction is shown in Figure 2.

U-1	I wonder which one is the fastest [looking at cards]
U-2	I think this one is fastest [touching a lion card], what do you think? [looking at robot]
R	I'm not sure about this, but I think the lion is the fastest animal [looking at cards]
U-1	Okay [moving the lion]
R	Now it looks better
U-2	Yeah… How about the zebra? [looking at robot]
R	I think the zebra is slower than the horse. What do you think? [looking at U-1]
U-1	I agree
U-2	I'm not sure, the zebra has to be fast to escape the lion…
R	mhm

Figure 2: *Example of interaction dialogue (the original language is Swedish)*

After the task was discussed for some time, a button is shown on the table that could be pressed to reveal the solution (see Figure 3 below). Furhat then commented on the solution, comparing it with his own belief (admitting mistakes or pointing out that they should have listened to him). After that, the players could play another round. Only the players can move the cards during the discussion.

Figure 3: *At the end of each game, the players receive feedback about the outcome of the game*

2.4. Data analysis

To investigate the accommodation, we choose 30 interactions from the original corpus [26], which includes recording of more than 390 interactions. We excluded interactions with child participants, as they will raise additional source of variability (age) and the interaction with less than two games played. We analyzed the first two games of each team thereby comparing the f_o distance between the speakers and the robot in the first two games.

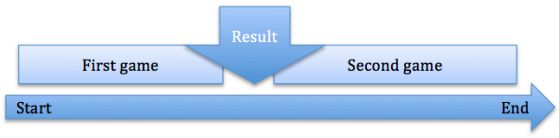

Figure 4: *Structure of the game. After finishing the first game, humans learned about their success (result) and then entered the second game.*

The fundamental frequency values of both the speakers and the robot were automatically extracted (To Pitch: 0, 75, 500 formula) using Praat software (version 6.0.43) [27]. For the current study we focus on the f_o values, which were measured separately for each utterance (see examples in Fig. 5)

To measure the f_o accommodation, we calculated the Euclidean distance in the speech of the two speakers, and between the robot and each of the human speakers during their game playing conversations. Then we measure the relation between those distances and (a) the duration of their play, (b) the game result.

We used linear mixed-effect models for f_o Euclidean distances based on the lmer() function in the lme4 package in the statistics software R [29]. The fixed effects are game result (winning or losing), and gender (same or mixed gender players).

The distance type (the distance between human-human or human-Robot) is considered to be a random factor for which a random intercept model was calculated. Significant interactions ($p < 0.05$) of the fixed effects were calculated by the Anova() function of the car package in R [30]. To examine if there are any significant differences between the amount of accommodation and distance type or interaction gender type, we calculated one-way ANOVA using R.

3. Results

3.1. Impressionistic analysis of f_o accommodation

Figure 5 shows f_o of the two human interaction partners (blue and green) and the robot voice as a function of time. A representative selection of graphs from the 30 interaction was picked. The dotted lines are polynomials fitted to the data of each speaker and represent the overall development of f_o over time. The vertical black line in each graph indicates the point at which game one ends and game two starts. The number at the top of the vertical line shows the result of game 1 (number of correct cards out of 5). The graphs are organized in female-female communication partners (left) mixed gender (center) and male-male (right). As can be seen, accommodation over time does not follow a systematic pattern. There is no obvious effect that f_o between the conversation partners approximates as a result of time or that f_o between humans and computers converges or diverges. According to a subjective visual analysis of the f_o curves, we distinguished the graphs into three categories, graphs where there is an apparent non-linear (random) relationship between the human conversation partners (top), a divergence between the partners (middle) and a synchrony between the partner (bottom). While such an analysis might be argued between different observers, it shows that the development of f_o over time is very complex and does not follow any obvious patterns. This seems true for both human-human and human-robot interactions. Such results are in line with previous literature [2] revealing that f_o accommodation between conversation partners is complex and

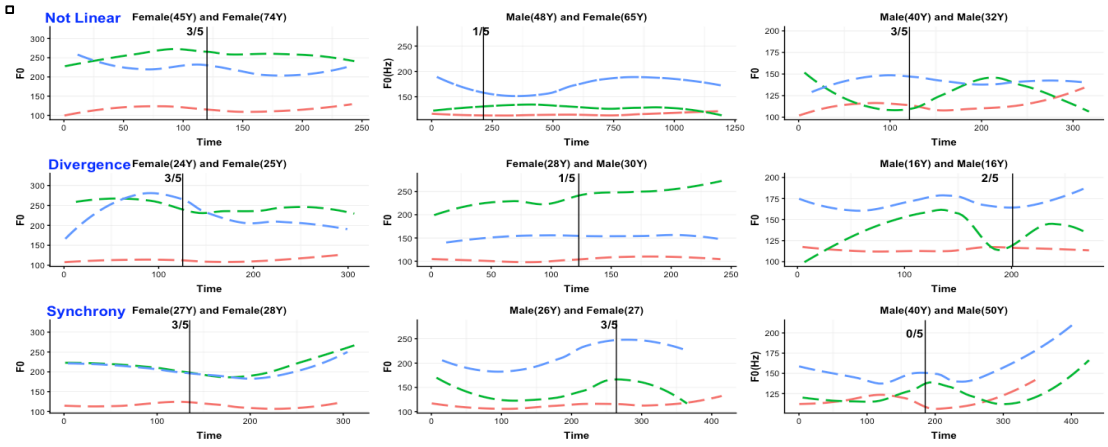

Figure 5: *Polynomials of f_o as a function of time for humans (green and blue) and robot voice (red) for a representative selection of interactions (see more details in the text)*

does not follow one specific pattern.

3.2. The relation between duration and accommodation

To measure the effects of the duration of the game on the accommodation between speakers we obtained the difference of Euclidian distance between the speakers in each pair (game 2- game 1) and plotted them as a function of time. Given that longer conversations allow more time to accommodate f_o between the partners, we would expect Euclidian distances to drift more towards negative values with longer duration of the conversation. This result, however, could not be obtained (Figure 6). The figure shows that the regression between Euclidian distance and time of conversation is very low (see Table 1) and Euclidian distances almost randomly vary around 0 at any point in time (Figure 6). It is possible that accommodation over time within each conversation would show different results but given the impressionistic observations obtained in Figure 5, there is no strong reason to assume that a clearer pattern can be found. In addition, calculating the f_o accommodation over time is very complex here as there are numerous interruptions of the f_o curve when speakers take turns and the polynomials in Figure 5 are only an approximation of the f_o curve. Another complex factor is that conversations are of very different duration, which must have an impact on the degree of f_o accommodation possible.

Table 1: *Linear regression results of (f_o Euclidean distance ~duration)*

Distance categories	R^2	Adjusted R^2
Two speakers	0.02855	-0.01193
Robot and speaker1	0.1598	0.1287
Robot and speaker 2	0.1064	0.07332

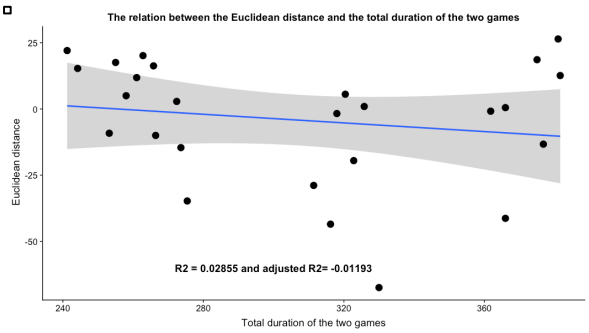

Figure 6: *The f_o difference of Euclidean distances of the played games in relation to the total duration*

3.2.1. The relation between accommodation and Gender composition of the teams

We examine the effect of gender of the players on the Euclidean distance during their interaction. Figure 7 shows the f_o distance difference in the first and second game. There are no visible differences in distance between the two games for different gender compositions. This is supported by an inferential analysis, for which no effect could be obtained (p=0.467).

3.2.2. The relation between game result and accommodation

Given that the interaction patterns were complex, we wanted to know whether there was any influence of the success in the first game on the f_o accommodation between the human interactants. For this, we measured the difference in f_o in all 30 groups of interactants during the first game and the second game. Figure 8 reveals the f_o distance between the different games (red = 1, blue = 2) between humans-humans and the robot voice for the different game outcomes (from 0 to 5 out of 5 cards correct). The figure shows that with higher numbers of correct cards the f_o distance decreased. To test this effect, we grouped conversations into two categories, (a) losing the game (0 to 2 cards out of five correct) and winning the game (3to 5 cards out of five correct). An inferential model showed that winning the game significantly influenced the amount of accommodation (t = -2.710, p = 0.006721).

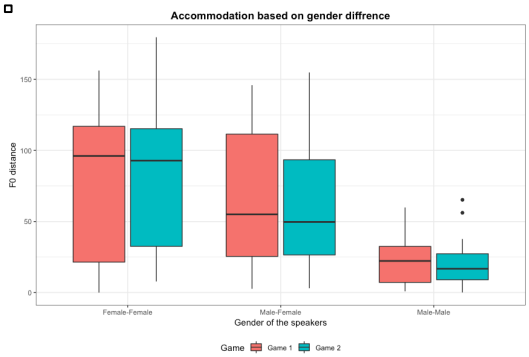

Figure 7: *The relation between the degree of accommodation and team gender composition*

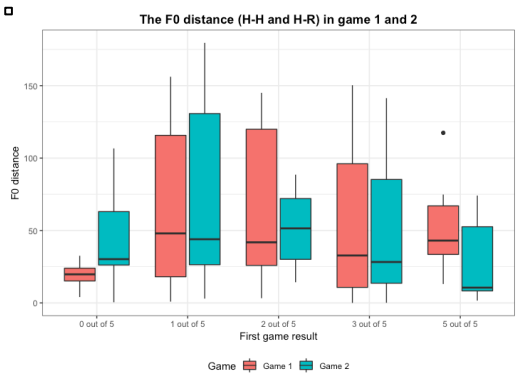

Figure 8: *The degree of accommodation and the game results (winning or losing)*

4. Conclusions

In the present paper, we investigated the influence of both the duration and the outcome of played card games on the degree of f_o accommodation in a semi-experimental situation. We analyzed 30 interactions where pairs of humans played a sequence of collaborative card sorting games with a robot. At the end of each game, they received feedback about the outcome of the game. We compared the f_o distance between their first two games. The findings suggest that accommodation between speakers is not necessarily a function of the duration of a conversation, but situational factors like winning the game, can influence speakers' convergence. We did not obtain support for the assumption that distance type (human-human and human-robot) and team gender composition (same or mixed gender interactions) affect the degree of accommodation.

5. References

[1] J. S. Pardo, "Measuring phonetic convergence in speech production." Frontiers in psychology, 4, 559. doi:10.3389/fpsyg.2013.00559, 2013

[2] J. S. Pardo, "On phonetic convergence during conversational interaction". Journal of acoustic society, Am. 119, 2382–2393 10.1121/1.2178720, 2006

[3] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability". Journal of Personality and Social Psychology, 32(5), 790, 1975.

[4] E Kurtić, J. Gorisch, "F0 accommodation and turn competition in overlapping talk", Journal of Phonetics, Volume 71, Pages 376-394, ISSN 0095-4470, 2018.

[5] J. Gorisch, B. Wells, and G. J. Brown, "Pitch contour matching and interactional alignment across turns: An acoustic investigation". Language and Speech, 55(1), 57–76, 2012.

[6] M. Zellers, A. Schweitzer, "An Investigation of Pitch Matching Across Adjacent Turns in a Corpus of Spontaneous German". Proc. Interspeech 2017, 2336-2340, DOI: 10.21437/Interspeech.2017-811, 2017

[7] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions". Proceedings of Interspeech 2011

[8] H. Giles, J. Coupland, and N. Coupland, "Accommodation theory: Communication, context, and consequence". In Contexts of accommodation: Developments in applied sociolinguistics: 1–68, 1991.

[9] R. Y. Bourhis, H. Giles, "The Language of Intergroup Distinctiveness. In Language, Ethnicity, and Intergroup Relations", ed. H. Giles. 119–135, 1977

[10] V. Delvaux, and A. Soquet, "The influence of ambient speech on adult speech productions through unintentional imitation". Phonetica 64(2-3): 145–173, 2007.

[11] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation". Journal of Phonetics 40(1): 177–189, 2012.

[12] N. Lewandowski, "Automaticity and consciousness in phonetic convergence". The Listening Talker: 71. Conference proceedings, 2012.

[13] S. E. Brennan, and H. Clark, "Conceptual pacts and lexical choice in conversation". Journal of Experimental Psychology: Learning, Memory and Cognition 22(6):1482– 1493, 1996.

[14] S. E. Brennan, "Conversation with and through computers. User modeling and user-adapted interaction",1(1):67–86,1991.

[15] J. Heath, "How automatic is phonetic convergence? Evidence from working memory". In Proceedings of the Linguistic Society of America, volume 2, page 35, 2017.

[16] J. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition, and contrast". In J. Bybee & P. Hopper, eds., Frequency and the Emergence of Linguistic Structure: 137–157, 2001

[17] M. Bane, P. Graff, and M. Sonderegger, M, "Longitudinal phonetic variation in a closed system," in Proc. of the 46th Chicago Ling. Soc., 2012

[18] J. S. Pardo, R. Gibbons, A. Suppes, R. Krauss , "Phonetic convergence in college roommates". Journal of phonetics 40, 190–197 10.1016/j.wocn.2011.10.001, 2012

[19] K. Drager, J. Hay, and A. Walker, "Pronounced rivalries: Attitudes and speech production". Te Reo, 53:27–53, 2010.

[20] U. Cohen Priva, and C. Sanker, "Distinct behaviors in convergence across measures". In Proceedings of the 40th Annual Meeting of the Cognitive Science Society, pages 1515–1520, 2018

[21] R. Levitan, A. Gravano, L. Willson, S. Ben˘us˘, J. Hirschberg, and A. Nenkova, "Acoustic- prosodic entrainment and social behavior". In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12, pages 11–19, Stroudsburg, PA, USA. Association for Computational Linguistics, 2012.

[22] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction". Speech Communication, 58, 11–34, 2014.

[23] D. N. Honorof and D. H. Whalen, "Perception of pitch location within a speaker's F0 range," Journal of the Acoustical Society of America, vol. 117, 2005.

[24] J. Bishop and P. Keating, "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex," The Journal of the Acoustical Society of America, vol. 132, no. 2, pp. 1100–1112, 2012.

[25] R. Levitan, S. Benus, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in AAAI 2015 Spring Symposium on Turn-taking and Coordination in Human- Machine Interaction, 2015.

[26] G. Skantze, M. Johansson, and J. Beskow, "Exploring Turn-taking Cues in Multi-party Human-robot Discussions about Objects". In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (pp. 67–74). https://doi.org/10.1145/2818346.2820749, 2015

[27] S. Al Moubayed, G. Skantze, and J. Beskow , "The furhat back-projected humanoid head-lip reading, gaze and multi-party interaction". International Journal of Humanoid Robotics. , 2013.

[28] P. Boersma, D. Weenink, "Praat: doing phonetics by computer" [Computer program]. Version 6.0.50, retrieved 31 March 2019 from http://www.praat.org/, 2019.

[29] D. Bates, M. Ma¨chler, B. Bolker and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4". Journal of Statistical Software, 67(1):1–48, 2015.

[30] J. Fox, and S. Weisberg, "An R Companion to Applied Regression". Sage, Thousand Oaks CA, Second ed., 2011.